ChinaXiv合作期刊 图书情报工作
第63卷第22期 2019年11月

基于网络结构和文本内容的群体画像构建方法研究*

■ 邱云飞 张伟竹

辽宁工程技术大学软件学院 葫芦岛 125105

摘要:[目的/意义]在基于社会网络的用户画像研究中,针对传统用户建模难以处理复杂网络关系,群体构建多基于内容,以及群体相似度低或紧密性差的问题,提出基于网络结构和文本内容的群体画像构建方法。[方法/过程]首先,采用卷积神经网络方法,融合网络结构和文本内容两方面特征将网络用户表示成空间向量,其次,在k-means算法基础上结合模块度计算方法,对空间向量进行聚类,然后,在爬取的中英文数据集上分别进行对比研究,最后,从中文数据集中选取1 000名重要性用户进行实例分析。[结果/结论]实验结果表明,该方法的密度值比基于内容的方法平均增加0.105,熵值比基于结构(含基于结构和内容)的方法平均减少0.955,实例分析进一步说明文中方法的可行性。

关键词:社会网络 网络关系 文本内容 深度学习 聚类算法 用户画像

分类号:TP391

DOI:10.13266/j.issn.0252-3116.2019.22.003

互联网时代,人们每天都会接触各种各样的网络平台,在这些平台上可以关注感兴趣的话题、浏览喜欢的内容,或者通过添加好友结交朋友、通过关注获得粉丝等,用户通过这些行为与他人建立联系,就如同在现实生活中一样,这种联系也会因朋友关系的变化而变化,最终形成网络。这种网络不单单指微博、知乎、Twitter等社交网络,还有知网、IEEE Xplore等引文网络,豆瓣、Digg等传播网络。在这些网络中,除了用户个人属性、发布内容、关注量等真实可见、直接可用的数据外,便是大量间接可得的网络关系,如关注关系、引用关系等。国内外学者基于网络关系进行网络表示学习、社区发现、用户画像等研究,其中用户画像成为近些年的一个热点研究,为个性化服务[1]、推荐系统[2]、精准营销[3]等带来巨大的应用价值,如何利用社会网络数据准确、全面、有效地刻画用户画像,成为众多学者们努力的方向。

1 相关研究

在早期的用户画像研究中,学者们根据用户的发布文本、关注内容、在线评论等对用户的社会属性、兴趣爱好、行为习惯、信誉度等进行预测分析。S. Alaoui等[4]从语义角度出发,根据用户发布的文本信息检测用户的购买意向,并结合用户的性别、年龄等社会属性信息进行产品推荐。W. X. Zhao等[5-6]根据用户在微博上的关注内容及发布文本,检测用户的购买意向,分别构建社交媒体用户画像和电商网站用户画像,并将两种画像进行映射和关联,实现基于社交媒体的电商产品推荐。单晓红等[7]以携程酒店为例,利用在线评论数据构建用户画像概念模型,对酒店用户特征进行刻画。余传明等[8]对股吧用户的发文内容进行深度表示学习,并结合用户的粉丝量、关注量、发帖量、评论量、吧龄等行为特征,提出一种行为—内容融合模型,以识别股吧用户是否属于噪声投资者。郭光明[9]通过对多源异构数据进行处理和分析,构建用户信誉画像并对用户信誉度进行预测评估。范晓玉等[10]融合个人主页、知网、基金网等多个数据源信息,提出融合多源异构数据的科研人员画像研究方法,并从基本属性、科研偏好、科研关系3个方面对科研人员进行画像分析。

随着研究工作的进展,人们发现在网络平台上,用户往往通过添加好友、相互关注、引用文本等方式与他人建立联系,鉴于此,学者们开始利用网络关系对网络

* 本文系国家自然科学基金青年科学基金项目"二向性反射分布函数的先验知识耦合式融合方法研究"(项目编号:61401185)研究成果之一。

作者简介:邱云飞(ORCID:0000-0002-2061-6617),副院长,教授,博士;张伟竹(ORCID:0000-0001-5450-8342),硕士研究生,通讯作者,E-mail:1426483346@qq.com。

收稿日期:2019-03-31 修回日期:2019-06-14 本文起止页码:21-30 本文责任编辑:王传清

中的用户建模,研究预测用户的标签信息,A. Mislove 等[11]根据社交网络中的关注关系构建网络拓扑结构并进行社区发现,采用聚类算法根据已知用户属性信息,预测未知用户属性,曹玖新等[12]利用微博关注关系拓扑结构,采用概率级联模型和机器学习方法对用户的转发行为进行预测,刘勘等[13]融合用户行为、用户发布内容以及社交关系等维度,采用随机森林算法对微博机器用户构建识别模型,徐志明等[14]将微博社会网络看做加权无向图,根据边的权值判断用户之间的相似性。虽然上述研究中融合了网络关系,但是在复杂网络中,机器学习方法通常具有局限性,尤其在无法提供大量的训练集时,传统方法的建模精度比较低。

为了简化对复杂关系的处理,学者们通过构建网络群体,从整体上对群体用户进行分析。林燕霞和谢湘生[15]根据社会认同理论定义微博主题,并对微博发布内容进行主题分类,最后采用多维标度法实现群体分类,张宏鑫等[16]通过构建移动终端日志数据主题模型,根据用户日志数据与主题的相关度进行聚类,熊伟等[17]利用 LDA 主题模型对网页内容划分主题,并根据用户行为信息分群画像,这些方法虽然避免了对复杂网络关系的处理,但仅基于用户文本内容进行聚类,虽然达到了内容上的相似,但群体的网络结构却不够紧密。为此,有些学者专门基于网络结构进行群体构建,以提高群体紧密度,如 V. D. Blondel 等[18]提出了根据社交网络图构建社群的方法,J. Leskovec 等[19]通过构造有向无权图对社交用户聚类成群,虽然基于结构的聚类方法可以提高群体的紧密度,但这些群体在内容或属性上未必相似。理想的网络群体构建方法不仅结构上紧密,而且内容上相似,因此,K. Steinhaeuser 和 N. V. Chawla[20]将节点属性作为网络图的边的权重,提出一种基于随机游走的群体构建方法,Y. Zhou 等[21]定义了一种结合结构和属性相似度的距离测量方法,将节点属性和边添加到图中以构建网络群体,Z. Xu 等[22]提出一种基于贝叶斯概率模型的图聚类方法,该方法从图结构和属性信息两方面建模,避免了对距离的计算,陈克寒等[23]根据用户微博内容相似度融合图摘要方法聚类建群,实现用户兴趣推荐的目的,吴树芳等[24]根据用户之间的关系,通过线性调和链入标签相似度和链出标签相似度,对用户的相似性进行度量。虽然这些研究在考虑结构的基础上尽量达到内容上的相似,但是群体的紧密度和相似度往往不可兼顾,即相似度高的群体紧密度低,紧密度高的群体相似度低,达不到理想的群体构建效果。

针对以上研究存在的问题,本研究利用卷积神经网络的方法,融合网络结构和文本内容两方面特征,对网络中的用户进行建模,通过将用户表示成空间向量来处理复杂的网络关系。将深度学习方法训练得到的用户表示向量进行聚类,在 k-means 算法基础上,利用结构模块度和内容模块度对聚类群体强度进行评估,构建网络群体,使群体内部在结构上更加紧密,在内容上更加相似,以实现较为理想的群体构建方法。

2 研究方法

2.1 网络用户表示

在各类网络平台上,用户都会通过加好友或者关注等行为与其他用户构建网络关系,网络中的用户一旦发生交互,用户的各种信息都有可能发生改变,可能根据交互用户的不同表现出不同方面的特点,也可能根据关注内容的不同与不同的用户交互。如在现实社交网络中有 A、B、C3 名学者,A 研究深度学习与数据挖掘,B 研究数据挖掘与自然语言处理,C 研究自然语言处理与用户画像,学者 B 与 A 可能由于数据挖掘相关的研究进行合作,学者 B 与 C 也可能一起研究自然语言处理,并且可以推断,学者 A 与 C 可能因为 B 达成合作,即采用深度学习的方法研究用户画像。针对这种复杂的网络关系,传统方法难以准确建模,为了能够较为准确地对网络用户进行表示,本研究采用深度学习的方法,通过深度训练对用户建模,除了考虑显式的网络关系以外,还根据文本的上下文语义信息来推断隐含的网络关系。

2.1.1 方法概述 在网络表示方法中,基于神经网络的方法通常在构建模型时设置一个损失函数,通过优化该函数找到更加合适的参数值,建立较为准确的模型,相比基于矩阵的方法往往运算速度快,执行效率高,精确度也会提高,因此本研究采用基于神经网络的方法将用户表示成空间向量形式,为了兼顾结构与内容两方面特征,分为结构表示向量和内容表示向量两部分。近几年,经典的基于神经网络的网络结构表示方法有 DeepWork、LINE 和 Nod2vec,其中,只有 LINE 方法生成的是上下文相关的节点表示,而后文由节点内容生成的"内容表示向量"也是上下文相关的,采用 LINE 方法更加有利于两者合并,另外,LINE 方法采用了一阶近邻和二阶近邻(一阶邻近是指直接相连的节点,二阶邻近是通过其他中介点相连的节点),这种表示也更加符合真实网络中用户之间的关系(用户之间通过直接关注建立联系或者由于关注了同一用户而建

立联系）。由于网络结构复杂且信息量较大，机器学习方法的处理效率比较低，而卷积神经网络方法在 GPU 配置环境下，处理速度非常快，另外，在对文本内容进行特征表示时，相比于 n-grams，卷积神经网络方法在表征高维特征时更具优势，因此，本研究的结构表示向量采用 J. Tang 等[25]提出的 LINE 模型来实现，内容表示向量采用卷积神经网络方法实现，最终将两种表示向量求和得到用户表示向量。

卷积神经网络架构共分为 3 层，将用户的文本内容和网络关系作为卷积神经网络的输入层，得到文本矩阵，将文本矩阵进行卷积、池化等操作，以此作为隐藏层，在输出层使用 softmax 函数对隐藏层结果进行向量归一化，得到单位向量，单位向量乘以文本矩阵得到内容表示向量。通过卷积神经网络生成内容表示向量的流程如图 1 所示：

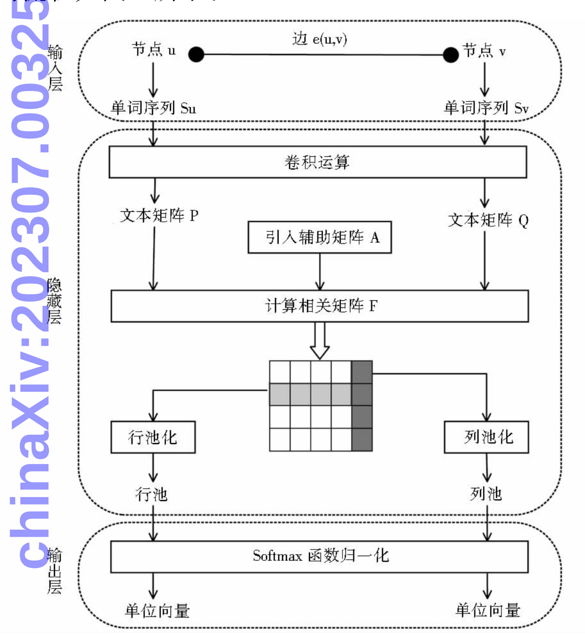


图 1　内容表示向量生成流程

2.1.2　方法实现

（1）过程描述。定义网络图 $G=(V,E,T)$，V 表示网络中的节点，$E\subseteq V\times V$ 是由节点构成的边，用来表示节点间的关系，T 表示节点的文本内容。由图 1 可见，以边 $e(u,v)\in E$ 上的两个节点 $u\in V,v\in V$，为例，利用卷积神经网络得到内容表示向量的具体过程如下：

第一，将两个节点的关注内容转换成单词序列 S_u、S_v，通过卷积层进行卷积运算分别生成 $P\in R^{d\times m}$ 和 $Q\in R^{d\times n}$ 两个文本矩阵，其中，m、n 分别表示 S_u、S_v 的长度，d 表示空间维度。

第二，引入辅助矩阵 $A\in R^{d\times d}$，计算相关矩阵 $F\in R^{m\times n}$ 为：

$$F=\tanh(P^{T}AQ) \tag{1}$$

第三，采用均值池对矩阵 F 的行和列分别进行行池化和列池化，得到 P、Q 的池化向量为：

$$p=[p_1,\cdots,p_m]^{T},\ q=[q_1,\cdots,q_n]^{T} \tag{2}$$

第四，利用 softmax 归一化函数将 P、Q 的池化向量转化成单位向量 a^p 和 a^q，q^p 的第 i 个元素表示为：

$$q_i^p=\frac{\exp(p_i)}{\sum_{j\in[1,m]}\exp(p_j)} \tag{3}$$

其中，$p_i=mean(F_{i1},\cdots,F_{in})$，表示 p 的第 i 个元素，即矩阵 F 每行池化的结果。a^q 计算同 a^p。

第五，单位向量与文本矩阵的乘积即为内容表示向量：

$$u_c=Pa^p,\ v_c=Qa^q \tag{4}$$

（2）损失函数。在统计学和机器学习中，损失函数通常是用来衡量错误和损失程度的函数，本研究将损失函数用于用户建模的评估，目的是最小化损失函数，提高建模精确度。由节点 u 预测交互对象为节点 v 的条件概率的相反数作为整个模型的损失函数，u 产生 v 的条件概率定义为：

$$\log(v|u)=\alpha\cdot\log(v_s|u_s)+\beta\cdot\log(v_s|u_c)+\beta\cdot\log(v_c|u_s)+\gamma\cdot\log(v_c|u_c) \tag{5}$$

$$\log(v_s|u_s)=\frac{\exp(u_s\cdot v_s)}{\sum_{x\in v}\exp(u_s\cdot x_s)} \tag{6}$$

其中，u_s、v_s、x_s 分别表示节点 u、v、x 的结构表示向量，α、β、γ 表示参数。

整个模型的损失函数定义为：

$$Loss=-\sum_{e\in E}\log(v|u) \tag{7}$$

2.2　网络群体构建

将模型训练得到的用户表示向量作为聚类算法的输入，提出基于网络结构和文本内容的群体构建方法（group construction method based on network structure and text content，GCNSTC），使得关系紧密且相似的用户聚类成群。采用 k-means 算法聚类，利用模块度对聚类结果进行评估并不断迭代更新，直到聚类结果不再改变。模块度由 M. E. J. Newman[26]提出，是一种用来衡量网络社区结构强度的度量值。为了使聚类效果更佳，从网络结构和文本内容两个角度定义模块度，将结构模块度与内容模块度加权求和作为最终的模块度评估指标。结构模块度与内容模块度的计算公式如下：

（1）结构模块度。利用传统的 Newman 模块度求解公式计算。

$$M_s = \sum_{(x,y) \in C} S(x,y)\delta(C_x, C_y) \quad (8)$$

$$S(x,y) = \frac{1}{2m}\left(A_{xy} - \frac{d_x \cdot d_y}{2m}\right) \quad (9)$$

其中,x、y 表示节点;C_x、C_y 表示群体;A_{xy} 表示当节点 x、y 相连时值为 1,否则为 0;d_x 表示节点 x 的度;m 表示网络中的总边数,$2m$ 表示整个网络的度;$\delta(C_x, C_y)$ 表示当 x、y 在同一个群体时值为 1,否则值为 0。

(2)内容模块度。将节点的文本内容转换成可度量的向量表示形式,利用余弦相似度计算内容模块度。

$$M_c = \sum_{(x,y) \in C} C(x,y) \quad (10)$$

$$C(x,y) = \cos(x_i, y_i) / \left(\sqrt{\sum_{i=1}^{d} x_i^2} \cdot \sqrt{\sum_{i=1}^{d} y_i^2}\right) \quad (11)$$

其中,d 表示文本表示向量的维度。

(3)将结构模块度和内容模块度加权求和得最终的模块度评估指标 M:

$$M = \omega M_s + (1 - \omega) M_c \quad (12)$$

其中,ω 为加权因子,$0 < \omega < 1$。

根据结构模块度和内容模块度,对网络用户进行聚类的算法见算法 1:

算法 1:基于结构_内容模块度的聚类算法

输入:用户表示向量,用户关注内容,初始群体个数 k
输出:群体个数 K,一组群体

```
调用 k-means 聚类算法,得到 k 个群体
Repeat
  for 节点 i do
    for 群体 j do
      if 节点 i 不在群体 j 中 do
        将节点 i 从其群体中移除并加入到群体 j
        计算将节点 i 加入后的模块度增量
      end if
    end for
  选择模块度增量最大的群体 j,将节点 i 移到群体 j,否则,节点 i 保留在原群体
  end for
until 群体不再发生变化
```

3 实验

3.1 数据获取及预处理

本文采用中英文两种数据集进行对比实验,用以说明文中方法的可行性以及对中英文数据集的普遍适用性,中文数据集采用从知乎爬取的用户数据,英文数据集则源于 A. K. Mccallum 等[27] 创建的英文引文网络 Cora。由于文章旨在融合网络结构和文本内容两方面特征对网络用户建模,因此,采集的数据包含文本数据和网络结构数据两部分。

3.1.1 知乎数据集 使用爬虫工具 ache 对知乎用户的关注话题内容及用户之间的关注关系进行爬取,ache 可以根据指定的搜索主题或属性内容,返回相关的搜索页面。在配置 ache 时设置爬取内容为“关注话题描述内容”,设置爬取的用户数量为 10 000,关注话题数小于等于 3。对于爬取的文本,首先采用 Python 正则表达式去除文本中含有的 html 标签内容,然后利用 jieba 分词对文本进行分词处理,使文本表示转化成词序列表示,然而中文文本处理常常遇到乱码问题,因此,在读取数据进行分词处理时采用 GBK 编码,进行分词处理后采用 utf8 编码存储数据。分词处理后文本中通常会有一些无效的词,比如,“这”“这个”“了”“什么”“呢”等,采用中文停用词表去除无效词,最终得到中文文本数据集。对于网络结构数据,每名用户都有一个唯一确定的编号 ID,当用户之间存在关注关系时,在数据集中存储两个用户的 ID 信息,如果没有关注关系,则不用存储。

利用 LDA[28] 主题模型对关注内容进行话题分类,知乎文本数据集(不完全列举)见表 1 所示。知乎结构数据集共含节点 10 000 个,边 43 894 条。

3.1.2 Cora 数据集 Cora 是一个经数据爬取后的引文网络,但是该网络数据比较大,本文从中筛选出有关“机器学习”的相关文章作为文本内容。对得到的文本进行预处理,使用 Python 类库 pyenchant 进行拼写检查更正,删除掉形如“liike”“lke”等拼写错误的词,然后进行英文文本分词。英文文本分词采用 Python 中 nltk 的 SnowballStemmer 类进行词干提取并使用 WordNetLemmatizer 类进行词形还原。另外,英文字母有大小写之分,如“Hello”和“hello”表示一个含义,但由于大小写的不同,往往被当做两个词,因此,需要将所有大写字母转换为小写,采用 Python 的 API 来实现。最后,英文文本中通常含有一些类似“of”“to”“an”“a”等无效词,通过引入英文停用词表去除无效词,得到实验所需英文文本数据集。对于网络结构数据,每篇文章都有一个编号 ID,根据筛选出的文章编号查找引用关系并存储。

从 Cora 引文网络中筛选出 2 277 篇机器学习相关论文,根据研究内容分成 7 个类别,Cora 文本数据集(不完全列举)见表 2。Cora 的结构数据集含节点为 2 277个,边为 5 214 条。

邱云飞，张伟竹. 基于网络结构和文本内容的群体画像构建方法研究[J]. 图书情报工作,2019,63(22):21 -30.

表 1　知乎文本数据集

用户	关注内容	话题 1	话题 2	话题 3
0	心理学 Psychology 一门 学科 注意 话题 心理 区别 研究 动物 心理现象......	心理学	设计	美食
1	电影 一种 视听 媒介 利用 胶卷 录像带 数位 媒体 影像 声音 捕捉 加......	电影	创业	运动健康
2	电影 一种 视听 媒介 利用 胶卷 录像带 数位 媒体 影像 声音 捕捉 加......	电影	互联网	旅行
3	生活 物质 生活 精神 生活 总称 物质 生活 生活 基本 需要 精神 生活......	生活	设计	阅读
4	电影 一种 视听 媒介 利用 胶卷 录像带 数位 媒体 影像 声音 捕捉 加......	电影	互联网	生活
5	最美 景色 永远 远方 再远 脚步 不出 心房 一种 社会 行为 使用 体育运......	旅行	运动健康	自然科学
6	心理学 Psychology 一门 学科 注意 话题 心理 区别 研究 动物 心理现象......	心理学	经济学	科学技术
7	教育 培养 新生 一代 准备 从事 社会 生活 整个 过程 人类 社会 生产......	教育	文学	物理学
8	电影 一种 视听 媒介 利用 胶卷 录像带 数位 媒体 影像 声音 捕捉 加......	电影	摄影	古典音乐
9	电影 一种 视听 媒介 利用 胶卷 录像带 数位 媒体 影像 声音 捕捉 加......	电影	互联网	旅行

表 2　Cora 文本数据集

编号	论文内容	类别
1	Graphical models enhance representational power probability models qualitative characterization properties This also leads greater efficiency terms......	2
2	Realtime Decision algorithms class incremental resourcebounded Horvitz anytime Dean algorithms evaluating influence diagrams We present test......	2
3	Speedup learning seeks improve computational efficiency problem solving experience In paper develop formal framework learning efficient problem......	3
4	This paper presents incremental concept learning approach identiflcation concepts high overall accuracy The main idea address concept overlap......	6
5	In previous paper SM showed finite automata could used define objective functions assessing quality alignment two sequences In paper show results......	0
6	Wahba Wang Gu Klein Klein introduced Smoothing Spline ANalysis VAriance SS ANOVA method data exponential families Based RKPACK fits SS......	0
7	This paper presents evolutionary approach incremental approach find learning rules several supervised learning tasks In evolutionary approach potential......	4
8	The overfit problem empirical learning utility problem explanationbased learning describe similar phenomenon degradation performance due increase amount......	5
9	We consider formal models learning noisy data Specifically focus learning probability approximately correct model defined Valiant Two widely studied......	3
10	We present decision tree based approach function approximation reinforcement learning We compare approach table lookup neural network function......	1

3.2　对比方法

本文采用 3 种群体聚类算法作为基线，分别从基于文本内容、基于网络结构以及基于结构和内容 3 个方面进行对比实验。

(1) K-means 算法：基于内容的聚类算法采用 K-means 作为基线，将文本内容转换成向量等形式，通过计算文本间的距离进行聚类。

(2) Louvain 算法：潘理等[29]指出社区发现算法只关注聚类结果在结构上的稠密性，而不考虑节点的属性信息，因此，基于结构的方法采用的是社区发现中的经典算法 Louvain 算法，该算法通过构造加权网络，利用节点之间的关系，将每个节点看做一个社区，不断计算将节点加入其邻居节点的模块度增益构建社群。

(3) SA-Cluster 算法：基于结构和内容的算法是采用 Y. Zhou 等[21]提出的 SA-Cluster 算法，将节点的属性信息添加到网络中建立增广网络，然后定义结构和属性相似度，并利用随机游走算法计算网络节点之间的距离，从而实现群体的构建。

3.3　评估指标

采用密度与熵两个评估指标对上述群体构建方法进行评估。密度主要反映群体内部成员之间关系的紧密程度，密度值越大，群体成员之间关系越紧密，密度的计算公式为：

$$D = \sum_{i=1}^{k} \frac{m_i}{m} \tag{13}$$

其中，k 表示群体个数，m 表示网络的总边数，m_i 表示群体 i 中的边数。

熵，原是热力学中用来度量体系中混乱程度的物理量，此处是用来反映群体内成员之间的相似度。如果社交网络中的节点加入某一群体后，导致熵值增大，则说明该节点的加入会引入额外的信息，因而，该节点与群体中其他节点的差异性较大，所以，熵值越小，混乱程度越低，群体成员之间越相似，熵的计算公式为：

$$E = -\sum_{i=1}^{k} \frac{n_i}{n} \cdot \sum_{j} p_{ij} \log(p_{ij}) \tag{14}$$

其中，n 表示网络的总节点数，n_i 表示群体 i 中的节点数，p_{ij} 表示群体 i 中具有类别 j 的节点所占百分比。

3.4　实验结果及分析

将本文方法同上述 3 种群体构建方法进行比较，分别在知乎数据集和 Cora 数据集上设置对比实验($\omega = 0.5$)，采用密度和熵两种评价指标进行评估，知乎数

据集的实验结果见表 3，Cora 数据集的实验结果见表 4。

表 3　知乎数据集实验结果

方法	密度	熵
K-means	0.33	0.00
Louvain	0.47	1.58
SA-Cluster	0.54	1.69
GCNSTC	0.39	0.64

表 4　Cora 数据集实验结果

方法	密度	熵
K-means	0.42	0.00
Louvain	0.66	1.70
SA-Cluster	0.71	1.83
GCNSTC	0.57	0.85

在知乎数据集上，GCNSTC 方法的密度高于 K-means 算法，低于 Louvain 和 SA-Cluster，说明 GCNSTC 方法在群体紧密度上优于 K-means 算法，但是比 Louvain 和 SA-Cluster 方法稍差些；通过熵值比较，K-means 算法的熵值为 0，群体内成员相似度最高，GCNSTC 方法比 Louvain 和 SA-Cluster 方法的熵值均小，群内成员的相似度比两者要好。

在 Cora 数据集中可见，GCNSTC 的密度值为 0.57，同样是高于 K-means 算法，低于 Louvain 和 SA-Cluster 方法，说明 GCNSTC 方法得到的群体在紧密度上较为理想；K-means 的熵值仍然是 4 个方法中最低的，在群体相似度上效果最好，但是 GCNSTC 方法的熵值跟 Louvain 和 SA-Cluster 方法相差 1/2 之多，在群体相似度上明显优于 Louvain 和 SA-Cluster 方法。

综上所述，通过在中英文两种数据集中对 GCNSTC 方法的聚类结果进行比较，得到相似的聚类效果，说明本文方法对中英文数据集具有普适性。通过密度值比较，虽然 GCNSTC 方法没有 Louvain 和 SA-Cluster 方法效果好，但优于基于内容的 K-means 聚类方法，平均密度值增加 0.105，因此就紧密度而言本文聚类方法较为理想。K-means 算法的熵值始终为 0，是因为 K-means 算法根据文本距离进行聚类，数据挖掘中通常用距离表示相异度，而相异度与相似度是一对相反的概念，因此距离越近内容越相似，所以在群体相似度上 K-means 算法都是上述方法中最好的，但是 GCNSTC 方法的熵值比 Louvain 和 SA-Cluster 方法平均减少 0.955，在群体相似度上，GCNSTC 方法优于基于结构的方法（包括基于结构和内容的方法）。

3.5　讨论

在群体相似度上，从上述实验结果结果可以看出，基于内容的 K-means 方法的熵值最低，群体的相似度最高，这是因为 K-means 算法只考虑了文本内容的相似性，仅根据文本内容之间的欧氏距离进行聚类，通过寻找距离聚类中心最近的文本找到相似群体；而 GCNSTC 方法则是将文本内容和网络结构进行融合，即在 K-means 聚类群体中引入了网络结构信息，根据信息论中熵的定义，熵反映了个体出现概率对整体信息出现的不确定性，在群体中引入低相似度的网络结构信息，导致整体相似的不确定性增加，所以 GCNSTC 方法的熵值高于 K-means 算法，但是相比 Louvain 和 SA-Cluster，GCNSTC 方法引入的额外信息较少，在一定程度上达到了次优。从紧密度来看，GCNSTC 方法的密度值虽然优于 K-means 算法，但相比其他两种方法较差，这是因为本文在结构表示向量的处理上，采用的 LINE 模型属于浅层神经网络模型，在与深度模型得到的内容表示向量合并后，深度模型得到的结果较浅层模型结果更精确，相对弱化了结构表示向量在整体用户表示向量中的存在，使得用户表示向量更加侧重于内容表示，因此，在实验结果中，GCNSTC 方法在结构紧密性上仅优于基于内容的 K-means 算法，后期将重点对该问题进行研究。

根据实验结果，将上述 4 种方法得到的群体紧密度和群体的相似度进行排名，从群体紧密度上看，SA-Cluster > Louvain > GCNSTC > K-means，GCNSTC 方法排名第三，仅优于 K-means 方法，从群体的相似度上看，K-means > GCNSTC > Louvain > SA-Cluster，GCNSTC 方法在 4 种方法中达到次优，与基于结构的方法比，达到最优，由此可见，本文所提方法在群体紧密度和群体相似度上均有改进，并且在群体相似度上改进效果更明显。为进一步说明此结论，将 GCNSTC 方法同另外 3 种方法的密度值和熵值进行比较，变化量见表 5。从表 5 中可以看出，虽然 GCNSTC 方法相对于 K-means 方法在密度上有所改进，但是与完全基于文本相似度的 K-means 方法相比，综合熵值之后整体情况不如 K-means 方法；而对于 SA-Cluster 和 Louvain 方法，在熵值上改进效果明显，整体而言优于 SA-Cluster 和 Louvain 方法。总之，本文所提方法在一定程度上是有效的，并且在熵值上效果更明显，构建的群体相似度较高，群体紧密性较差。

表5 密度值和熵值比较结果

方法	知乎数据集			Cora数据集		
	密度	熵	综合	密度	熵	综合
K-means	↑0.06	↓0.64	↓0.58	↑0.15	↓0.85	↓0.70
Louvain	↓0.08	↑0.94	↑0.86	↓0.09	↑0.85	↑0.76
SA-Cluster	↓0.15	↑1.05	↑0.90	↓0.14	↑0.98	↑0.84

4 实例分析

统计中文数据集中每名用户的粉丝数量并降序排序,前50名中用户编号ID均在1 000以内,前100名中,只有两名用户的编号不在1 000以内,前200名中约有89%的用户ID在1 000以内,并且编号小于1 000的所有用户及其粉丝用户大约占据了整个网络的80%,本文将这1 000名用户设为整个网络的重要性用户。本文选取这1 000名用户,从群体内容的相似性和结构的紧密性两个角度,对构建的群体进行分析,每名用户含有1至3个关注话题。利用文中聚类方法将用户聚成10个群体,本文规定群体中每个话题的关注量占该群体话题关注总量的10%以上,即为主要关注话题,经统计,每个群体的成员数、话题关注总量、主要关注话题等信息汇总见表6。

表6 群体及其关注话题信息

群体	成员数	关注话题总量	主要关注话题(关注量)	主要话题占比(%)	平均主要话题占比(%)
1	55	164	互联网(54) 创业(41) 电影(21)	70.7	23.6
2	31	62	文学(9)	14.5	14.5
3	78	212	电影(52) 美食(41)	43.9	22
4	135	371	心理学(106) 经济学(87)	52	26
5	122	346	创业(213)	61.5	61.5
6	263	781	互联网(250) 电影(181) 旅行(124)	71.1	23.7
7	185	549	生活(139) 电影(119) 旅行(92)音乐(64)	75.4	18.9
8	71	179	心理学(19)	10.6	10.6
9	30	77	摄影(31) 电影(19)	64.9	32.5
10	30	73	电影(8)	11	11

从主要关注话题总量上看,有6个群体的主要话题关注量占比超过50%,表明这些群体的用户关注话题与群体整体关注内容大体吻合,用户之间在内容上相似度较高;虽然群体3的主要关注话题占关注话题总量的43.9%,但其主要关注话题只有两个,从整体上看,群体用户比较相似;群体2、8、10相比其他群体而言,聚类效果不太理想,群体主要关注话题占比比较低,说明群体内成员关注话题比较多,并且每个关注话题的关注量比较低(低于10%),群体成员之间差异性较大。从平均主要关注话题量上看,群体5的聚类效果最好,其主要关注话题只有"创业",且关注量达61.5%,说明该用户群体主要关注的是创业内容,个别用户关注的是其他话题,但关注量不多;虽然群体7的主要关注话题量高达75.4%,但平均主要关注话题量却只有18.9%,可见其群体整体主要关注内容较多,相似度较低。

针对上述10个群体,统计每名用户的关注用户,按其数量的降序排序。由于有些群体的成员数较少,而有些群体的关注用户有多个,因此,选择关注量大于10且排名前3的用户汇总见表7。从表7中可以看出,群体2、8、10的成员关注同一用户的数量超过整个群体的一半,这些成员以关注的用户为中心形成子群体,子群体占整个群体的50%之多,因此,整个群体结构相对比较紧凑;群体3和群体9相比前3个群体,紧密度降低,群体中含有一些未列举的小的子群体,虽然每个子群体不大,但总体上却削弱了群体的紧密度;剩余5个群体,成员关注同一用户的关注量都低于50%,从最大的子群体看,与群体3和群体9相差不大,但是这5个群体含有一些相对较大的子群体,对整个群体

表7 群体内用户及其关注用户信息

群体	关注用户(关注量)	占群体百分比(%)
1	414(27) 113(19) 2(15)	49.1 34.5 27.3
2	65(24)	77.4
3	48(32)	41
4	20(53) 167(45)	39.3 33.3
5	95(47) 113(32) 252(14)	38.5 26.2 11.5
6	2(105) 110(74) 9(53)	39.9 28.1 20.2
7	28(81) 640(56) 26(38)	43.8 30.3 20.5
8	20(49)	69
9	27(12)	40
10	48(18) 9(11)	60 36.7

的影响比较大，如群体 4，列举的两个子群体相差不大，从整体上看，相当于将群体 4 分成两个子群体，虽然子群体内部结构紧密，但整体性却比较差，另外，这 5 个群体中还有一些未列举的小群体，因此，整体结构相对比较松散。

通过上述实例分析，群体 2、8、10 在内容的相似性上虽然比较差，但在结构上相对比较紧凑，其他群体在内容上比较相似，但结构上却相对松散，从而也进一步说明，基于网络结构和用户文本内容进行群体构建，很难在结构紧密度和内容相似度上同时取得最优值。采用本文方法得到的群体，在熵值上提升较大，更加倾向于内容的相似性，因此，在 1 000 名重要性用户分析中，构建的 10 个群体，有 7 个群体在内容上比较相似，虽然结构上相对比较松散，但仍然可以将群体分成稍大的结构紧密的子群体，另外，有些用户同时关注了多个用户，在关注量上存在重叠，使得整个群体在结构上紧密度低降。

5　群体画像分析及研究意义

以群体 6 为例，进一步分析群体画像研究的意义。

（1）通过整体分析，可以进行用户分析、产品推荐、行业发展趋势预测等。经统计群体 6 主要关注的是互联网、电影、旅行、生活 4 个方面的内容，从电影、旅行、生活可以看出该群体用户比较注重生活品质，关注娱乐消遣；从互联网、电影可以看出群体中部分用户可能从事互联网影视行业。可以向该群体推荐一些新上映的影视作品，或者推荐一些旅游景区，分享一些旅游攻略等。该群体中互联网的关注量最高，其次是电影、旅行，在“互联网 +”时代，可以对电影和旅游业的发展情况进行预测，如在互联网 + 电影行业，除了可以网上影评、互联网购票，还可以在移动终端上随时观看影视作品；又如，在互联网 + 旅游行业，可以在社交网站创建一些社区，分享一些游记等，激发人们的旅游兴趣，可以设计一些旅游攻略软件，帮助人们对旅游景区、衣食住行等快速做出决策，还可以通过携程等电商平台，帮助人们购买车票、机票以及预定酒店等。

（2）通过内容角度分析，可以进行群体消息推送或群体推荐、朋友推荐等。统计群体 6 中每个话题的关注用户，部分结果见表 8。同一话题群体中，用户的关注内容相同，可以实现群体消息推送或群体推荐，如若有一些新上映的或评分比较高的电影，可以向电影群体推荐，也可以推送一些有关电影业的消息；当一张新专辑发布需要推广时，可以对音乐群体中的用户进行群体推荐，还可以细化音乐类型进行精准推荐。此外，通过协同过滤算法可以实现朋友推荐、个性化服务等，如若用户 21 想了解一些心理学知识，可以将用户 554 推荐给用户 21，因为两者都关注了电影、互联网话题，可能在心理知识方面有相似的诉求，可以将用户 554 了解的心理学知识推荐给用户 21；若用户 662 想通过互联网创业，可以将用户 662 的设计内容推荐给用户 129，因为两者都关注互联网和经济，说明两者可能从事互联网经济相关的职业，并且用户 129 还关注了电子商务，对电子商务领域有一定的了解，用户 662 的设计产品可以通过用户 129 发布于电子商务平台，实现网络经济活动。

表 8　群体 6 每个话题的关注用户部分结果

关注话题	用户											
	15	21	27	129	164	220	374	472	554	662	699	999
电影	√	√	√					√	√			√
电子商务				√								
互联网	√	√		√	√	√	√	√	√	√	√	√
经济				√			√	√		√		
旅行					√		√					√
设计										√	√	
生活		√			√	√					√	
心理学			√						√			
音乐	√		√			√						

（3）通过结构角度分析，可以进行用户分析、朋友推荐等。经统计，群体 6 中约有 40% 的用户关注用户 2，统计用户 2 及其关注用户的关注话题，其中，大部分关注电影、互联网、生活、旅行 4 个方面，其次是对音乐、经济学、创业、心理学及运动与健康方面的关注。从以用户 2 为关注对象的群体中可以看出，这些用户

可能更加关注娱乐消遣、注重生活品质,以看电影、旅游、听音乐等为主;论其职业,这些用户可能从事互联网影视,或者是互联网创业者,可能对商业或投资比较感兴趣;他们中的一些人会关注一些心理学知识,注意工作压力的排遣,调节心理压力等。统计用户 2 的部分关注用户及关注话题见表 9,以用户 2 为中介,可以实现朋友推荐,如,若用户 92 想了解音乐方面的信息,可以将用户 15、404 或 944 推荐给用户 92,因为用户 2 与 4 名用户都具有关注关系,因此可以通过用户 2 使用户 92 与 3 名用户联系,其中优先推荐用户 15 和 404,因为用户 92 与两者的共同偏好较多;若用户 404 想要选择一款汽车自驾旅游,可以推荐用户 426,因为两者都关注旅行话题,且用户 426 关注汽车话题,可能对适合自驾出游的汽车有所了解,能给予一些建议,因为两名用户都关注用户 2,可以通过用户 2 成为好友。

表 9　用户 2 的部分关注用户及关注话题

话题 1	话题 2	话题 3	用户 2 的关注用户									
			2	15	92	302	359	404	426	532	944	965
电影	互联网	旅行	√		√							
		音乐		√								
	旅行	音乐						√				
	生活	心理学								√		
互联网	创业	设计										√
	旅行	生活				√						
	游戏	Android					√					
旅行	创业	汽车							√			
	生活	音乐									√	

6　结语

本文根据现实网络中用户之间的复杂关系,采用深度学习的方法将网络用户进行空间向量表示,从网络结构和文本内容两个方面对网络用户建模并进行聚类,利用模块度对聚类群体强度进行衡量,通过不断迭代提高聚类效果。为验证文中方法的可行性,采用中英文两种数据集,与 3 种不同类型的聚类算法进行对比实验。实验结果表明,本文方法对中英文数据集具有普适性,并且该方法的密度值比基于内容的方法平均增加 0.105,熵值比基于结构的方法(含基于结构和内容的方法)平均减少 0.955,同时提高了群体紧密性与群体相似度。本文对实验结果进行讨论,阐述了在密度和熵值上均未达到最优的原因,通过综合分析密度和熵的变化量,说明本文方法在熵值的效果更明显,构建的群体更加侧重于相似度,而群体紧密性略差。最后,本文选择中文数据集中的 1 000 名重要性用户,从文本内容相似性和网络结构紧密性两个角度进行验证分析,分析表明,基于网络结构和文本内容进行群体构建,很难在结构紧密度和内容相似度上同时取得最优值,但文中方法在熵值上提升较大,实例分析中约有 7/10 的群体在内容上比较相似。通过对 1 000 名重要性用户进行分析,阐明了群体画像研究对产品推荐、行业预测、个性化服务、消息推送、用户分析以及朋友推荐等具有重要意义。

目前,针对社会网络进行群体画像研究,由于群体构建文本特征单一,群体画像刻画比较泛化,另外,对群体结构的研究尚存在不足,因此,在后期研究中除了融入多特征数据来分析预测用户的其他属性标签外,还将重点进行群体结构的研究。

参考文献:

[1] 何娟. 基于用户个人及群体画像相结合的图书个性化推荐应用研究[J]. 情报理论与实践,2019,42(1):129－133,160.

[2] ZHAO W X, WANG J, HE Y, et al. Mining product adopter information from online reviews for improving product recommendation[J]. ACM transactions on knowledge discovery from data, 2016, 10(3):1－23.

[3] 刘海, 卢慧, 阮金花, 等. 基于"用户画像"挖掘的精准营销细分模型研究[J]. 丝绸,2015,52(12):37－42,47.

[4] ALAOUI S, AJHOUN R, IDRISSI Y E B E, et al. Semantic approach for the building of user profile for recommender system[C]// Global summit on computer & information technology. Sousse: IEEE, 2016:114－119.

[5] ZHAO W X, GUO Y, HE Y, et al. We know what you want to buy: a demographic-based system for product recommendation on microblogs[C]// ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2014: 1935－1944.

[6] ZHAO W X, LI S, HE Y, et al. Exploring demographic information in social media for product recommendation[J]. Knowledge

and information systems, 2016, 49(1):61 – 89.

[7] 单晓红, 张晓月, 刘晓燕. 基于在线评论的用户画像研究——以携程酒店为例[J]. 情报理论与实践,2018,41(4):99 – 104,149.

[8] 余传明, 田鑫, 郭亚静, 等. 基于行为 – 内容融合模型的用户画像研究[J]. 图书情报工作,2018,62(13):54 – 63.

[9] 郭光明. 基于社交大数据的用户信用画像方法研究[D]. 合肥:中国科学技术大学,2017.

[10] 范晓玉, 窦永香, 赵捧未, 等. 融合多源数据的科研人员画像构建方法研究[J]. 图书情报工作,2018,62(15):31 – 40.

[11] MISLOVE A, VISWANATH B, GUMMADI K P, et al. You are who you know: inferring user profiles in online social networks [C]// ACM international conference on web search and data mining. New York: ACM,2010:251 – 260.

[12] 曹玖新, 吴江林, 石伟, 等. 新浪微博网信息传播分析与预测[J]. 计算机学报,2014,37(4):779 – 790.

[13] 刘勘, 袁蕴英, 刘萍. 基于随机森林分类的微博机器用户识别研究[J]. 北京大学学报(自然科学版),2015,51(2):289 – 300.

[14] 徐志明, 李栋, 刘挺, 等. 微博用户的相似性度量及其应用[J]. 计算机学报,2014,37(1):207 – 218.

[15] 林燕霞, 谢湘生. 基于社会认同理论的微博群体用户画像[J]. 情报理论与实践,2018,41(3):142 – 148.

[16] 张宏鑫, 盛风帆, 徐沛原, 等. 基于移动终端日志数据的人群特征可视化[J]. 软件学报,2016,27(5):1174 – 1187.

[17] 熊伟, 杭波, 李兵, 等. 一种集成用户画像与内容的服务重定向方法[J]. 小型微型计算机系统,2017,38(12):2762 – 2765.

[18] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008(10):10008 – 10019.

[19] LESKOVEC J, LANG K J, MAHONEY M W. Empirical comparison of algorithms for network community detection[C]// ACM international conference on World Wide Web. Raleigh:ACM, 2010: 631 – 640.

[20] STEINHAEUSER K, CHAWLA N V. Identifying and evaluating community structure in complex networks[J]. Pattern recognition letters, 2010, 31(5):413 – 421.

[21] ZHOU Y, CHENG H, YU J X. Graph clustering based on structural/attribute similarities[J]. Proceedings of the VLDB endowment, 2009, 2(1):718 – 729.

[22] XU Z, KE Y, WANG Y, et al. A model-based approach to attributed graph clustering[C]// ACM SIGMOD international conference on management of data. Scottsdale:ACM, 2012:505 – 516.

[23] 陈克寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报,2013,36(2):349 – 359.

[24] 吴树芳,徐建民,武晓波. 融合用户标签和关系的微博用户相似性度量[J]. 情报杂志,2014,33(12):170 – 173,126.

[25] TANG J, QU M, WANG M, et al. LINE: large-scale information network embedding[C]// International conference on World Wide Web. Florence:WWW, 2015:1067 – 1077.

[26] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical review e statistics nonlinear soft matter physics, 2003, 69(6):066133.

[27] MCCALLUM A K, NIGAM K, RENNIE J, et al. Automating the construction of internet portals with machine learning[J]. Information retrieval journal,2000, 3(2):127 – 163.

[28] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1):993 – 1022.

[29] 潘理, 吴鹏, 黄丹华. 在线社交网络群体发现研究进展[J]. 电子与信息学报,2017,39(9):2097 – 2107.

作者贡献说明:
邱云飞:提出论文研究思路,指导修改论文;
张伟竹:设计研究方案并进行实验,撰写论文。

Study for the Construction Method of Group Profile Based on Network Structure and Text Content

Qiu Yunfei Zhang Weizhu
Liaoning Technical University, Huludao 125105

Abstract: [**Purpose/significance**] In the study of user profile based on social network, aiming at the problems that traditional user modeling is difficult to deal with the complex network relationship, group construction is mostly based on content, and the group is low similarity or poor tightness, a construction method of group profile based on network structure and text content is proposed. [**Method/process**] Firstly, using the convolutional neural network method, the network structure and the text content are combined to represent the network user as a space vector. Secondly, based on the k-means algorithm, the modularity calculation method is combined to cluster the space vector. In the crawled Chinese and English datasets, a comparative study is conducted. Finally, 1000 important users are selected from the Chinese dataset for instance analysis. [**Result/conclusion**] The experimental results show that the density value of this method is increased by 0.105 compared with the content-based method, and the entropy value decreases by 0.955 on average compared with the structure-based (including structure-based and content-based) method. The instance analysis further illustrates the feasibility of the proposed method.

Keywords: social network network relationship text content deep learning clustering algorithm user profile